Prediction of wine quality using machine learning techniques

Erestina Gjeka, University of Donja Gorica, Podgorica, Montenegro erestina.gjeka@udg.edu.me

Abstract — Climate change has affected every sector of nature, especially healthcare in recent years. These changes have affected the vineyards but also the characteristics of the wine. In this research project, two natural factors were taken into account, temperature and annual precipitation. At times when machine learning had not yet been discovered, each process was very complicated and time-consuming. Therefore, machine learning is a very smart move to get fast and accurate results. Pearson correlation coefficient was used to come to a conclusion.

Keywords: Temperature · Precipitation · Wine · Pearson correlation coefficient.

1 INTRODUCTION

Problems due to climate change have led to debates on many topics, research, and even this research paper. Because of all this, there was a need to do various studies to determine the impact of temperature and precipitation on wine quality. This is very important because it gives us information on how to use temperature and precipitation to get the type of wine we want. This research paper will deal specifically with the analysis of temperature and precipitation as the main factors influencing the quality of wine. Two types of wine were considered: white and red, the two most commonly used types. Italy is known for its wine and the variety of grapes it produces; that is why the information for the research paper is based on this country; however, the same method of analysis can be used for any other country. In order to study the data on each wine in more detail and more precision, three hypotheses were made. Depending on these hypotheses, what truly affects the quality of wine can be seen. A detailed analysis was made for each of them.

- Hypothesis 1: The amount of alcohol in the wine depends on the temperature.

- Hypothesis 2: Precipitation affects the acidity of wine.

– Hypothesis 3: Precipitation affects the pH of the wine.

2 PEARSON CORRELATION COEFFICIENT

The Pearson correlation coefficient is often referred to only as the correlation coefficient. Below are its characteristics and how it can be used. The correlation coefficient allows clear monitoring of the relationships that the two variables have with each other. This coefficient gives results that show how strong the relationship between two variables is [1].

The formula used to calculate the Pearson correlation coefficient is:

$$r_{XY} = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i) (\sum_{i=1}^{n} y_i)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$
(1)

 r_{XY} = Pearson correlation coefficient

n = number of samples

x =first variable

y = second variable.

In statistics, the correlation coefficient is very important because, in addition to strength, it also gives the direction in which the two variables move. The direction can take two forms: positive or negative. Positive means that if one variable grows, so does another, and if one decreases, so does the other. If a negative result is obtained, it means that one variable grows while the other simultaneously decreases. It is necessary to be careful when choosing variables because it is not good to analyze two variables that have little to do with each other. That's because, even if we get results that show a high impact between two variables, we cannot say exactly that this is a completely correct solution if we have selected variables incorrectly. The choices made for the analysis to be further studied used logic that shows that temperatures and precipitation really affect wine production. However, this should be checked to determine the accuracy or eventual inaccuracy of the claim [2].

As noted earlier, the correlation coefficient gives us two pieces of information about the variables: the direction and strength of their connection. The direction has already been processed; we are left with the strength of the connection between the two variables. If:

- $-r_{XY} = +1$; this is the perfect positive link
- $r_{XY} \ge 0.75$; this is a pretty strong positive link.
- $-0.25 \le r_{XY} \le 0.75$, this is a moderate positive link.

If r_{XY} has the value greater than 0 and less than 0.25, then that connection is positive but weak. If $r_{XY} = 0$, it means that if one variable increases or decreases, the other variable

©The author(s). Open access is under the terms of the Creative Commons Attribution Non-Commercial No Derivatives 4.0 International Licence.

does not change. If r_{XY} has the value greater than -0.25 and less than 0, then that connection is negative but weak.

 $-r_{XY} = -1$; this is the perfect negative link

- $r_{XY} \leq -0.75$; this is a pretty strong negative link.

- -0.75 $\leq r_{XY} \leq$ -0.25, this is a moderate negative link [1].

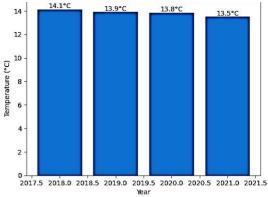
The correlation between the variables is displayed using graphics known as heat maps. Heat maps allow data to be visualized via matrices. It is very useful, especially in cases where links between multiple variables are sought, not just two. The matrix is created in such a way that it allows us to notice key data with one glance and to study the strength of the connections that the variables have with each other. Variables are displayed through columns and rows, while the numbers we get in the matrix are indicators of the strength of the relationship between each two variables [3].

3 ANALYSIS OF PRECIPITATION AND TEMPERATURE IMPACTS ON WINE CHARACTERISTICS

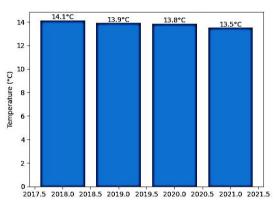
Data analysis is a process that must be carefully considered, both in data selection and in further work. During this analysis, we will not only get a link between the characteristics of wine and temperature or rain, but we will also get much more than that thanks to heat maps.

The country to which all the data relates is Italy, because it is one of the largest wine producers in the world. There are many wine-producing cities, but among the highest quality are cities like Sicily, Tuscany, Piedmont, and many others. Italy is a country where climatic conditions differ, which means that each zone in which grapes can be grown differs from another zone. Even within the same zone, there may be small variations in temperature and precipitation. Since Italy is not a small country, the average temperature and precipitation values for the whole country have been taken into account. The average wine characteristics in some famous wine-producing cities are also determined [4]. The aim of the analysis is to link the effects of climate characteristics on wine production in order to act on its production in the best possible way. Data were collected through online websites, not only for temperatures and precipitation but also for wine. All data were analyzed using the Pearson correlation coefficient and presented via different graphs. This analysis aims to conclude how much effect temperature and precipitation have on wine and what characteristics of wine are affected by them.

The initial step is to find the data. The rain data in Italy used in the project is from this website [5], and the temperature data were taken from the website [6]. Data collected and grouped is always desirable to be displayed using charts. The main goal is to compare the annual level of data from 2018 to 2021. Initially, precipitation data will be analyzed, and years will be identified with the largest and smallest amounts of precipitation. Once these results are obtained, the same process will be repeated for temperatures. The second step will be to present the characteristics of the wine. Two types of wine, red and white, were considered in this analysis. Graphs will allow the characteristics of white wine and red wine to be compared. The three main characteristics to be considered during this analysis are alcohol, acidity, and pH.



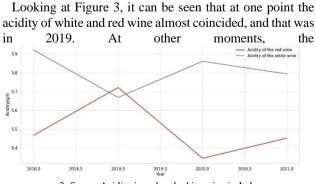
1. figure: Average precipitation value in Italy for the years 2018-2021.



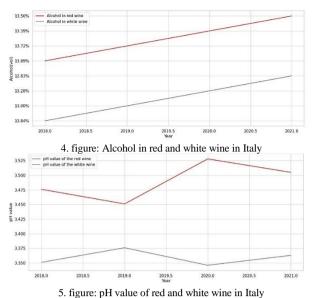
2. figure: Average temperature value in Italy for the years 2018-2021

Figure 1 shows a precipitation graph for Italy over the past 4 years. Displaying data in this way allows information to be processed faster. The largest precipitation was recorded in 2019, with 890.8mm. This shows the average amount of precipitation in 2019, expressed in millimeters. On the other hand, it was noted that the precipitation in 2021 was lower compared to previous years, at 730.6mm. In Figure 2, the temperature table for Italy over the past four years is shown.

It can be observed that the highest temperature was recorded in 2018, at 14.1 °C. The lowest temperature compared to others was that of 2021, with 13.5 °C. Some graphs related only to the characteristics of the wine will be displayed. The three charts below will compare the properties of white and red wine.



3. figure: Acidity in red and white wine in Italy



acidity values of red and white wine do not approach each other; on the contrary, the acidity of red wine is almost always lower than the acidity of white wine.

It can be said that, in general, for these years, the acidity of red wine is lower than that of white wine. The most interesting part of this graph is the difference between white and red wine, as it will be further used for analysis.

In Figure 4, which is shown below, looks as if there is a linear correlation between variables; when one grows or falls, the other does the same. However, a pattern isn't really shown, so no general statement can be drawn.

From Figure 5, it can be concluded that the pH value of red and white wine is never the same. Red wine always has a higher pH value than white wine.

The main graphics are shown with basic data for climate and wine characteristics. All the elements needed to form a Pearson correlation coefficient were found.

3.1 Pearson correlation coefficient calculated for precipitation and properties of wine

Table 1 shows the assessment using the Pearson correlation coefficient that uses formula (1). In the Figure 6, the code that has been used to form the Table 1 below is shown.

whitewine = pd.read_csv('/kaggle/input/white-wine-and-precipitation-in-italywhitewine.head()

redwine = pd.read_csv('/kaggle/input/red-wine-and-precipitation-in-italy-2021 redwine.head()

data_merged = pd.merge(whitewine, redwine, on="mm.")
data_merged.head()

corr_matrix = data_merged.corr(method="pearson")
corr_matrix.head(10)

6. figure: The code for calculating the Pearson correlation coefficient for precipitation and wine characteristics

	mm.	alcohol_x	acidity_x	pH_x	alcohol_y	acidity_y	pH_y
mm.	1.000	-0.316	-0.827	0.720	0.168	0.842	0.654
alcohol_x	0.316	1.000	0.760	0.621	0.600	-0.281	0.082
acidity_x	0.827	0.760	1.000	0.924	0.083	-0.812	0.518
pH_x	0.720	-0.621	-0.924	1.000	0.228	0.913	0.730
alcohol_y	0.168	0.600	0.083	0.228	1.000	0.511	0.808
acidity_y	0.842	-0.281	-0.812	0.913	0.511	1.000	0.919
pH_y	0.654	-0.082	0.518	0.730	-0.808	-0.919	1.000

1. table: The results shown were calculated using the Pearson correlation coefficient for precipitation and wine characteristics.

Table 1 gives the results of the Pearson correlation coefficient for precipitation. Each variable that has a suffix x belongs to white wine, while those that have a suffix y belong to red wine. This table shows the result calculated by the Pearson correlation coefficient, which establishes a connection between every two elements. Not only by analyzing the relationship between temperature and precipitation with the characteristics of wine but also by analyzing the link between the properties of red and white wine.

Through the Pearson correlation coefficient, we can confirm or refute the hypotheses formed at the beginning of this analysis. As mentioned earlier, all results obtained are between -1 and 1, with -1 showing a perfectly negative link where one variable decreases while the other grows and 1 showing a perfect positive bond where one variable grows or decreases while the other also increases or decreases.

Data with values between -0.25 and 0.25 cannot provide reliable information that can be confirmed with certainty because the link between them is not strong enough, which means that their connection is weak. On the other hand, results above 0.25 and below -0.25, but especially those above 0.75 and below -0.75, can provide information with greater certainty as their connection is stronger.

The obtained data via the Pearson correlation coefficient will be displayed in the form of heat maps to make it easier to read. The heat map will display the results of the Pearson correlation coefficient for each type of wine separately. As seen in Table 1, there is also an analysis of the relationship between the characteristics of both wines, but this link is not significant for further analysis. Therefore, the results will be displayed separately in the white and red wine heat map.

"Seaborn" and "Matplotlib" have been two packages that have helped to earn the desired results (heat maps); using them, the coding has become easier and better visualized. On the other side, "font_scale" and "fig_size" can be used to arrange the photo how we desire it to be. In all of the heat maps, "font_scale" has been equal to 1.1, and "fig_size" has been equal to (8, 8).

```
import seaborn as sns
import matplotlib.pyplot as plt
corr_matrix1 = whitewine.corr(method="pearson")
print(corr_matrix1)
sns.set(font_scale=1.1)
plt.figure(figsize=(8, 8))
sns.heatmap(corr_matrix1, annot=True, cmap="coolwarm", square=True, vmin=-1, vmax=1)
```

plt.show()

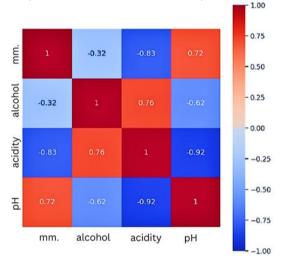
^{7.} figure: The code that has been used to gain the heat map for precipitation and characteristics of the white wine

The first heat map is generated using the code shown in Figure 7. This heat map depicts the properties of white wine with precipitation.

The code shown in Figure 8, was used to generate a heat map of red wine properties and precipitation.

In Figure 8, a graphic display represents the Pearson correlation coefficient calculated only for white wine and precipitation. In the same way, a heat map is obtained for red wine which is shown in Figure 10. The analysis will be performed in a way that takes into account only data that has a strong link.

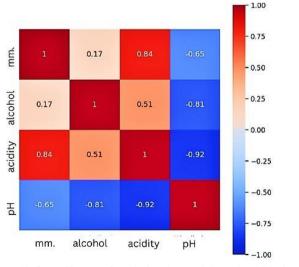
The heat map shown in Figure 8 shows that the following factors must be taken into account for the impact of precipitation on white wine properties: acidity and pH value. The acidity has a strong negative correlation, while the pH value has a moderate positive correlation.



8. figure: Heat map for white wine characteristics and precipitation

```
import seaborn as sns
import matplotlib.pyplot as plt
corr_matrix2 = redwine.corr(method="pearson")
print(corr_matrix2)
sns.set(font_scale=1.1)
plt.figure(figsize=(8, 8))
sns.heatmap(corr_matrix2, annot=True, cmap="coolwarm", square=True, vmin=-1, vmax=1)
plt.show()
```

9. figure: The code that has been used to gain the heat map for precipitation and characteristics of the red wine



10. figure: Heat map for red wine characteristics and precipitation

3.1.1 Analysis of the impact of precipitation on the acidity of wine.

White wine has an acidity value of -0.83. This suggests that acidity and precipitation in wine are inversely proportionate. When one number rises, the other falls. When there is less rain, the acidity is higher; when there is more rain, the acidity is lower.

According to the book [7] the relationship between wine and rainfall (precipitation) may be reinforced. The quantity of acid in wine is frequently responsible for its sourness. Wine, as is known, is produced from grapes, and if the grapes are in their best condition, then the wine will also taste better. Grapes ripen at their own pace; nevertheless, a perfectly ripe grape equivalent to grapes picked the previous season can never be achieved, even if they are cut at the same time and location but in consecutive years. The amount of sour wine is influenced by the age of the grapes. According to the same book [7] "too much rain during flowering time in late spring/early summer can lead to poor fruit set and, later on, increased fungal diseases and low yields."

According to this book, apple acid levels rise in colder temperatures while falling in warmer ones. It is possible to define chilly weather situations as those that are inherently distinguished by rain. Of course, it should be kept in mind that not all grape varieties are affected negatively by this association. Because of the positive correlation between red wine acidity and precipitation, which means that when precipitation grows, acidity also grows, and when precipitation decreases, acidity also decreases, it is important to delve into greater depth. White wine, on the other hand, has the opposite impact.

Red wine and white wine are manufactured using distinct techniques, and typically, white wine is made from grapes that are stronger than those used to make red wine. Additionally, it uses grapes that are less developed than those used to make red wine. The graph in Figure 3 supports the assertion that white wine has higher acidity than red wine. The book [8] says: "Red grapes tend to have a longer growing season and thicker skins, which make them more resilient to changes in weather conditions... For white grapes, excessive rainfall during the ripening period can lead to a reduction in acidity and a loss of aromatic intensity. White grapes tend to ripen earlier than red grapes and have a shorter growing season, so they are more vulnerable to changes in weather conditions.", which supports the idea that has been validated. While the heat map in Figure 8 indicates that the ratio between acidity in white wine and precipitation is strong and negative, the heat map in the other Figure 10 indicates that the relationship between acidity in red wine and precipitation is strong and positive (value of +0.84). We may infer that *hypothesis 2* is verified as true based on these two manuals, the books [3] and [4], and based on the study done using Pearson's correlation coefficient.

3.1.2 Analysis of the impact of precipitation on the pH value of wine.

In Figures 8 and 10, it is shown that there is a moderate relationship between pH and precipitation. Figure 8 displays the findings of the Pearson's correlation coefficient, which demonstrates a positive relationship between precipitation and the pH of white wine. Precipitation and red wine's pH have a negative connection, just like previously. This implies that while one number increases, the other decreases. A more thorough investigation is also necessary in this final instance, when white wine and red wine should be separated.

According to Article [9], it is stated that "Usually, if a wine has a high acid level, it will have a low pH.". Indirect analysis can be used to determine the link between pH and rainfall levels. It has been shown that acidity has an impact on pH value, and since pH is a phrase that indicates how much acidity is present in a substance, this enables an understanding of the relationship between these two terms. Additionally, precipitation in this instance is related to both acidity and pH. It was said that white wine's acidity and rainfall had an inverse relationship, meaning that as rainfall rises, acidity falls and vice versa.

The acidity of red wine, on the other hand, has a positive relationship with rainfall, which implies that these two values rise or fall simultaneously. As was already mentioned, there is a positive correlation between rainfall and white wine's pH whereas a negative correlation exists between red wine's pH and rainfall. Analyzing the relationship between pH and acidity for both wines is sufficient.

The acidity of white wine has a negative relationship with pH, as seen by the heat map in Figure 8. This means that when the pH value falls, acidity increases, which makes sense because the lower the pH value, the higher the acidity. In a prior analysis, it was determined that as the acidity of white wine increases, the levels of precipitation decrease since there is a negative correlation between the two. This leads to the conclusion that, when calculated based on acidity, pH value decreases when rainfall is reduced. This demonstrates once more that precipitation and pH value are positively correlated in white wine.

A red wine analysis will now be addressed. The acidity of red wine has a positive link with rainfall, as previously demonstrated, but this can also be observed in Figure 10 where the heat map is shown. This means that acidity and rainfall both rise and reduce at the same time. As a result, there is a direct link. Red wine, on the other hand, has a negative association with acidity. This demonstrates that as the pH value decreases, acidity increases in the same manner that it does in white wine, as this is a general concept that always applies. The examination of the Pearson correlation coefficient calculation in the heat map in Figure 10 leads to the conclusion that as precipitation falls, the pH value lowers. Because it has been indirectly proven that precipitation influences the pH of both wines, *hypothesis 3* is accepted.

3.2 Pirson correlation coefficient calculated for temperature and wine properties

The data in the figures were analyzed using the Pearson correlation coefficient, and the table is the result of that analysis with the aid of the following code:

<pre>whitewine = pd.read_csv('/kaggle/input/white-wine-and-temperature-in-italy-in-' whitewine.head()</pre>
<pre>redwine = pd.read_csv('/kaggle/input/red-wine-and-temperature-in-italy-in-2021- redwine.head()</pre>
<pre>data_merged = pd.merge(whitewine, redwine, on="celcius") data_merged.head(15)</pre>
<pre>corr_matrix = data_merged.corr(method="pearson") corr_matrix.head(15)</pre>
11 figure: The code for calculating the Pearson correlation coefficient

11. figure: The code for calculating the Pearson correlation coefficient for temperature and wine characteristics

2. table: The results shown were calculated using the Pearson correlation coefficient for temperature and wine characteristics.

	celcius	alcohol_x	acidity_x	pH_x	alco hol_ y	acidity_y	pH_y
celcius	1.000	0.867	0.351	0.267	0.70 2	0.149	0.424
alcohol_x	0.867	1.000	0.760	0.621	0.60 0	-0.281	0.082
acidity_x	0.351	0.760	1.000	0.924	0.08 3	-0.812	0.518
pH_x	-0.267	-0.621	-0.924	1.000	0.22 8	0.913	0.730
alcohol_y	0.702	0.600	0.083	0.228	1.00 0	0.511	0.808
acidity_y	0.149	-0.281	-0.812	0.913	0.51 1	1.000	0.919
pH_y	-0.424	-0.082	0.518	- 0.730	0.80 8	-0.919	1.000

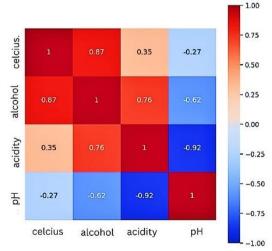
Through Table 2, we can see which of the data has a positive, negative, strong, or weak relative. Data that have a continuation of x refer to white wine, while data that have a continuation of y refer to red wine.

The identical procedure that was used for precipitation will be used for temperature. Formula (1) has been used to evaluate data relationships. Table 2 shows Pearson's correlation coefficient for the temperature and properties of white and red wine in the same table. Through Table 2, it is possible to analyze in detail what correlation exists between temperature and white or red wine regarding their acidity, alcohol, or pH values.

As highlighted in Table 1, we will divide this information into two correlation maps, where the Pearson correlation coefficient will be calculated, especially for red and white wine. As previously done when analyzing data on wine in relation to precipitation, only strong correlations are important. As has already been said, the data is easier to analyze using a heat map. The following are two heat maps where the published results were taken from Table 2. Using the coloring of the heat map, it is easily noticed which are strong and weaker distances.

The first heat map is generated using the code shown in Figure 12.

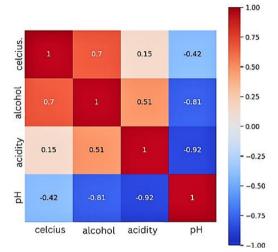




13. figure: Heat map for white wine characteristics and temperature

import seaborn as sns import matplotlib.pyplot as plt corr_matriz2 = redwine.corr(method="pearson") print(corr_matrix2) sns.set(font_scale=1.1) plt.figure(figsize=(8, 8)) sns.heatmap(corr_matrix2, annot=True, cmap="coolwarm", square=True, vmin=-1, vmax=1) plt.show()

14. figure: The code that has been used to gain the heat map for temperature and characteristics of the red wine



15. figure: Heat map for red wine characteristics and temperature

The same way as before, first the code will be shown, and then the output of that code, which will be the second heat map. The code in Figure 14 was used to generate a heat map of red wine properties and temperature.

In the first heat map shown in Figure 13, which represents a correlation map for Pearson's correlation coefficient between white wine characteristics and temperature, we see that only alcohol has a strong correlation with temperature. This means that for white wine, we will only analyze the relationship between temperature and alcohol. As for the second heat map in Figure 15, showing the correlation map for Pearson's correlation coefficient calculated for the relationship between the red wine and temperature properties, we can notice that only alcohol has a strong correlation with temperature. This means that we will only analyze that correlation here.

3.2.1 Analysis of the impact of temperature on alcohol in wine

The first heat map in Figure 13, displays a correlation map using Pearson's correlation coefficient between alcohol and temperature in white wine, which is 0.87. This suggests that in white wine, there is a positive relationship between alcohol and temperature. The quantity of alcohol in wine increases as the temperature rises. However, no mention is made of temperatures during fermentation, but rather of climatic temperatures during grape growing. However, because alcohol is temperature sensitive, further evidence is required to prove this association. This information is derived from a link [10], which states that temperature and climatic change affect wine and the level of alcohol in wine. This link says: "Warmer temperatures mean lower acidity and higher sugars in the grapes. Sugar converts to alcohol, so the end effect is a powerful acidity in the wine and higher alcohol levels." As a conclusion, wine is obtained with a higher amount of alcohol and a lower amount of acid. This proves that the temperature has an impact on the amount of alcohol in white wine.

Similarly, it has been established that temperature influences the alcohol content in red wine. The second heat map in Figure 15 displays the findings of the Pearson correlation coefficient between red wine qualities and temperature. Alcohol and temperature have a high reciprocal association, according to Pearson's correlation coefficient. This connection has a value of 0.7, indicating that alcohol and temperature have a significantly strong relationship.

Also, red wine as well as white wine have the same relationship in terms of alcohol and temperature. In the same way as with white wine, data were collected regarding this link [10], and based on Pearson's correlation coefficient, it can be concluded that temperature affects alcohol for red wine as well. As a conclusion, hypothesis 1 is confirmed as correct.

4 CONCLUSION

Many pieces of information from various books and websites were gathered during this analysis. In addition to the literature, the Pearson correlation coefficient and correlation maps were used, which showed the results of the coefficient. Temperature data were collected from the [5] and [6] websites, while wine data were collected from the website [11].

Initially, it was considered that the general notion that precipitation and temperature impact both white and red wine qualities had a foundation. The broad theory is then subdivided into three prosthetic hypothesis. According to this study, temperature and precipitation should be employed in favor of growing vines and creating the desired type of wine. Certain sites and methods of grape cultivation are essential.

The influence of climate change on vineyards should be considered every year and more, as climate change is becoming a severe issue. As a result, their influence on vineyards cannot be overlooked.

REFERENCES

- [1] Turney, Sh. (2020). Pearson Correlation Coefficient (r) | Guide and Examples.
- [2] Frost, J. Interpreting Correlation Coefficients, Statistics by Jim.
- [3] Kumar, A. (2022). Correlation Concepts, Matrix & Heatmap using Seaborn .
- [4] MacNeil, K. (2015). The Wine Bible. 2nd edn. Workman Publishing, New York, pages: 34, 601-609.
- [5] https://tradingeconomics.com/italy/precipitation (Accessed: 21.03.2023)
- [6] https://tradingeconomics.com/italy/temperature (Accessed: 21.03.2023)
- [7] Buglas J., A. (2022). An Introduction to Viticulture, Winemaking and Wine: From Vineyard to Wine Glass. Cambridge Scholars Publishing, UK, page 66.
- [8] Robinson, J. (2015). The Oxford Companion to Wine, Oxford University Press.
- Hale, N. (2022). What is acidity in wine?, Wine Enthusiast. <u>https://www.winemag.com/2019/06/19/what-is-acidity-in-wine/</u>
- [10] https://www.thirtyfifty.co.uk/spotlight-climatechange.asp#:~:text= Warmer\% 20temperatures\% 20mean\% 20lower\% 20acidity,feel\% 20lighter\ % 20in\% 20the\% 20mouth. (Accessed: 01.04.2023)
- [11] https://www.superiore.de/ (Accessed: 20.03.2023)